Main Manuscript for

DNA ligases of *Prochlorococcus marinus*: an evolutionary exception to the rules of replication

Erik Hjerde[2], Ashleigh Maguren[1], Elizabeth Rzoska-Smith[1], Bronwyn Kirby[1] and Adele Williamson[1, 2]*

[1] School of Science, University of Waikato, Hamilton 3240, New Zealand

[2] Department of Chemistry, UiT The Arctic University of Norway, Tromsø, N-9037, Norway

* Adele Williamson

Email: adelew@waikato.ac.nz

ORCID: 0000-0001-8139-1071

Author Contributions

AW designed the research; AM, ERS. BK and AW carried out the research; EH and AW analyzed the data; AW wrote the paper

Competing Interest Statement: Authors have no competing interests

This file includes:

Main Text
Figure 1

Word Count: = 1778

Reference Count: = 15

Abstract

DNA ligases, essential enzymes which re-join the backbone of DNA come in two structurally-distinct isoforms, NAD-dependent and ATP-dependent, which differ in cofactor usage. The present view is that all bacteria exclusively use NAD-dependent DNA ligases for DNA replication, while archaea and eukaryotes use ATP-dependent DNA ligases. Some bacteria also possess auxiliary ATP-dependent DNA ligases; however, these are only employed for specialist DNA repair processes. Here we show that in the genomes of high-light strains of the marine cyanobacterium *Prochlorococcocus marinus*, an ATP-dependent DNA ligase has replaced the NAD-dependent form, overturning the present paradigm of a clear evolutionary split in ligase usage. Genes encoding partial NAD-dependent DNA ligases are found on mobile regions in high-light genomes and lack domains required for catalytic function. This constitutes the first reported example of a bacterium that relies on an ATP-dependent DNA ligase for DNA replication and recommends *P. marinus* as a model to investigate the evolutionary origins of these essential DNA-processing enzymes.

Introduction

DNA ligases, enzymes that join breaks in the phosphodiester backbone of double-stranded DNA, are essential for DNA replication and repair in all organisms. They are classified as ATP-dependent (AD-ligases) or NAD-dependent (ND-ligases) on the basis of the adenylating cofactor used during catalysis (1). AD- and ND-ligases have different taxonomic distributions among cellular organisms and possess distinct sequence and structural features. ND-ligases are almost entirely limited to eubacteria where they are essential for replication, whereas eukaryotes and archaea use AD-ligases. This key difference in cellular replication machinery represents a central delineation between bacterial and archaeal/eukaryotic cell lineages (2); however the reason why bacteria preferentially use of NAD-ligases remains a long-standing question in the field (3). Previously, we described how strains of the marine cyanobacterium *Prochlorococcus marinus* encode multiple AD-ligases in their genomes, yet lack the other subunits associated with known AD-ligase-dependent repair pathways (4). *P. marinus* ecotypes are broadly classified as high-light which reside in the UV-damaging nutrient-poor upper ocean and low-light which experience less UV, and have higher nutrient access (5). Low-light strains have the smallest genomes of any free-living organisms (1.66 -1.75 MB) yet despite this minimization, they encode up to three predicted AD-ligases with different domain compositions (Fig 1A). To gain further insight, we have used a comparative genomics approach to survey the diversity of DNA ligases among sequenced isolates of *P. marinus*, which indicates a role for these AD-ligases in DNA replication.

Results and Discussion

Our comparative analyses of 41 *P. marinus* genomes (13 complete, 28 scaffolds) found that in high-light strains of *P. marinus*, the genes for the 'essential' replicative ND-ligases are severely truncated encoding polypeptides of <250 amino acids (Fig 1B). These truncated ND-ligases (hereafter ND-Lig tr) lack the oligonucleotide-binding domain and DNA-binding elements (BRCT and Zn-finger) necessary for activity (6). All high-light *P. marinus* strains harbor at least two AD-ligases (AD-Lig P and AD-Lig B) and MIT9312 has a third AD-ligase AD-Lig W (Fig C). In contrast, low-light strains possess full-length ND-ligases with all domains required for DNA ligase activity (ND-Lig fl). No low-light strains have an AD-ligase, however members of low-light clade LLI, which occupies an intermediate evolutionary position between ecotypes have both full-length and truncated forms of ND-ligase.

Our extensive attempts to recombinantly produce ND-lig tr were unsuccessful as the expressed protein was insoluble; however, given the lack of essential catalytic domains, it is highly unlikely that these truncated proteins are active. The lack of an intact ND-ligase gene indicates that in the high-light strains, all replicative and repair functions must be carried out by one or more of its AD-

ligases. To determine where these AD-ligases are located the high-light *P. marinus* genome, we carried out whole-genome alignments of representatives from each phylogenetic clade. Strikingly, we found that in high-light strains, AD-Lig P, occupies an identical genomic position to the replicative ND-Lig fl of the low-light ecotypes (Fig 1D). In all cases, the DNA ligase (either ND-lig fl or AD-lig P) is sandwiched between two small conserved hypothetical proteins and downstream of the replicative recombinase RecA. This analogous sequence context indicates that AD-Lig P has genetically substituted the ND-ligase in high-light strains, and suggests it may also functionally replace it in DNA replication. As we have previously described, AD-Lig P is active on singly nicked or cohesive breaks and prefers Mg as a divalent cation (4, 7). We further confirmed the predicted cofactor preferences of ND-Lig fl and AD-Lig P as NAD and ATP by assaying purified recombinant protein (Fig 1E). We also demonstrated through comparison of specific activities, that AD-Lig P is more than 10 fold as effective at sealing singly-nicked double-stranded DNA substrates relative to the low-light ND-Lig fl under equivalent conditions.
Genome rearrangements have played a key role in the evolution of *P. marinus* ecotypes (8), therefore we investigated whether any of these DNA ligases are located within mobile regions. No bacteriophages or genomic islands are predicted in regions containing ND-Lig fl, AD-Lig P, or AD-Lig B, however AD-Lig W which is only found in only a few strains is located in a known high-light genomic island (9). The truncated ND-ligases are also located in these genomic islands and their positions in *P. marinus* genomes vary considerably. No synteny in genes adjacent to ND-lig tr was found between different high-light strains, or with genes flanking the full-length ND-ligases of low-light ecotypes. In addition, both ND-lig tr genes of low-light LLI strain NATL2A (which also has a full-length ND-ligase) are located in an incomplete prophage, suggesting that these truncated ligases are not simply pseudogenes of a former replicative ligase, but may be horizontally transferred between strains. A phylogenetic tree of the ND-ligases supports this idea, showing that the truncated isoform forms a separate and more heterogenous clade (Fig 1F). Based on this, we propose that the ND ligases were truncated and duplicated in plastic regions of a sub-set of low-light strains, prior to replacement of the functional full-length ND-ligase by an ATP-dependent form in high-light strains.

Conclusions

To summarize, our genetic and biochemical analyses show that: 1) All high-light strains of *P. marinus* lack a gene encoding a complete ND-ligase gene, meaning this essential role must be fulfilled by one of its AD-ligases; 2) The AD-ligase AD-Lig P is substituted in the equivalent genetic context of the ND-ligase suggesting this isoform is responsible for replicative ligase activity; 3) The high-light AD-Lig P is a *bona fide* ATP-dependent DNA ligase with >10x higher specific activity relative to the low-light ND-ligase. The finding that an ATP-ligase must carry out both repair and replication in high-light *P. marinus* overturns the present paradigm of a clear evolutionary split between AD- and ND-ligase usage. This evidence that high-light strains of *P. marinus* are an evolutionary exception to the rules of replication recommends the various ecotypes as model systems to address the question of why most bacteria exclusively employ ND-ligases for DNA replication. As advances are being made towards genetic manipulation of *P. marinus* it may soon be possible to address this aspect directly (10).

Materials and Methods

Complete genomes for *P. marinus* were downloaded from ProPortal and NCBI (11) (see S1 for strain details and accession numbers). DNA ligases were identified in all genomes by searching for the nucleotidyltransferase domain of ND-ligases (PF01653) and AD-ligases (PF01068). Full-length coding sequences were retrieved and additional ligase-associated domains detected using Pfam as described previously (4). Whole genome alignments were constructed using progressiveMAUVE with default settings (12). Bacteriophage and genomic islands were identified using Alien_hunter (13) and Phaster (14). Phylogenetic trees used the Maximum likelihood method with Tamura-Nei model (DNA gyrase B) or using the JTT model (ND-ligases)

and 1000 bootstrap replicates. Recombinant ND-Lig fl (MIT9211, WP_012196358) was expressed and purified as described for AD-Lig P (7), and both proteins were assayed described therein.

Acknowledgments

We thank the Marsden Fund of New Zealand (18-UOW-034, AW) and Research Council Norway (244247, AW and EH) for funding. ERS is supported by a University of Waikato study award.

References

1. Williamson A & Leiros H-KS (2020) Structural insight into DNA joining: from conserved mechanisms to diverse scaffolds. *Nucleic Acids Research*.
2. Makarova KS & Koonin EV (2013) Archaeology of eukaryotic DNA replication. *Cold Spring Harb Perspect Biol* 5(11):a012963.
3. Doherty AJ & Suh SW (2000) Structural and mechanistic conservation in DNA ligases. *Nucleic Acids Research* 28(21):4051-4058.
4. Williamson A, Hjerde E, & Kahlke T (2016) Analysis of the distribution and evolution of the ATP-dependent DNA ligases of bacteria delineates a distinct phylogenetic group 'Lig E'. *Mol Microbiol* 99(2):274-290.
5. Rocap G*, et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424(6952):1042-1047.
6. Wang LK, Nair PA, & Shuman S (2008) Structure-guided mutational analysis of the OB, HhH, and BRCT domains of Escherichia coli DNA ligase. *J Biol Chem* 283(34):23343-23352.
7. Williamson A & Leiros HS (2019) Structural intermediates of a DNA-ligase complex illuminate the role of the catalytic metal ion and mechanism of phosphodiester bond formation. *Nucleic Acids Res* 47(14):7147-7162.
8. Biller SJ, Berube PM, Lindell D, & Chisholm SW (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nature reviews. Microbiology* 13(1):13-27.
9. Coleman ML*, et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311(5768):1768-1770.
10. Laurenceau R*, et al.* (2020) Toward a genetic system in the marine cyanobacterium *Prochlorococcus*. *Access Microbiology*.
11. Kelly L, Huang KH, Ding H, & Chisholm SW (2012) ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Research* 40(Database issue):D632-D640.
12. Darling AE, Mau B, & Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.
13. Vernikos GS & Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 22(18):2196-2203.
14. Arndt D*, et al.* (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 44(W1):W16-21.
15. Muhling M (2012) On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environ Microbiol* 14(3):567-579.

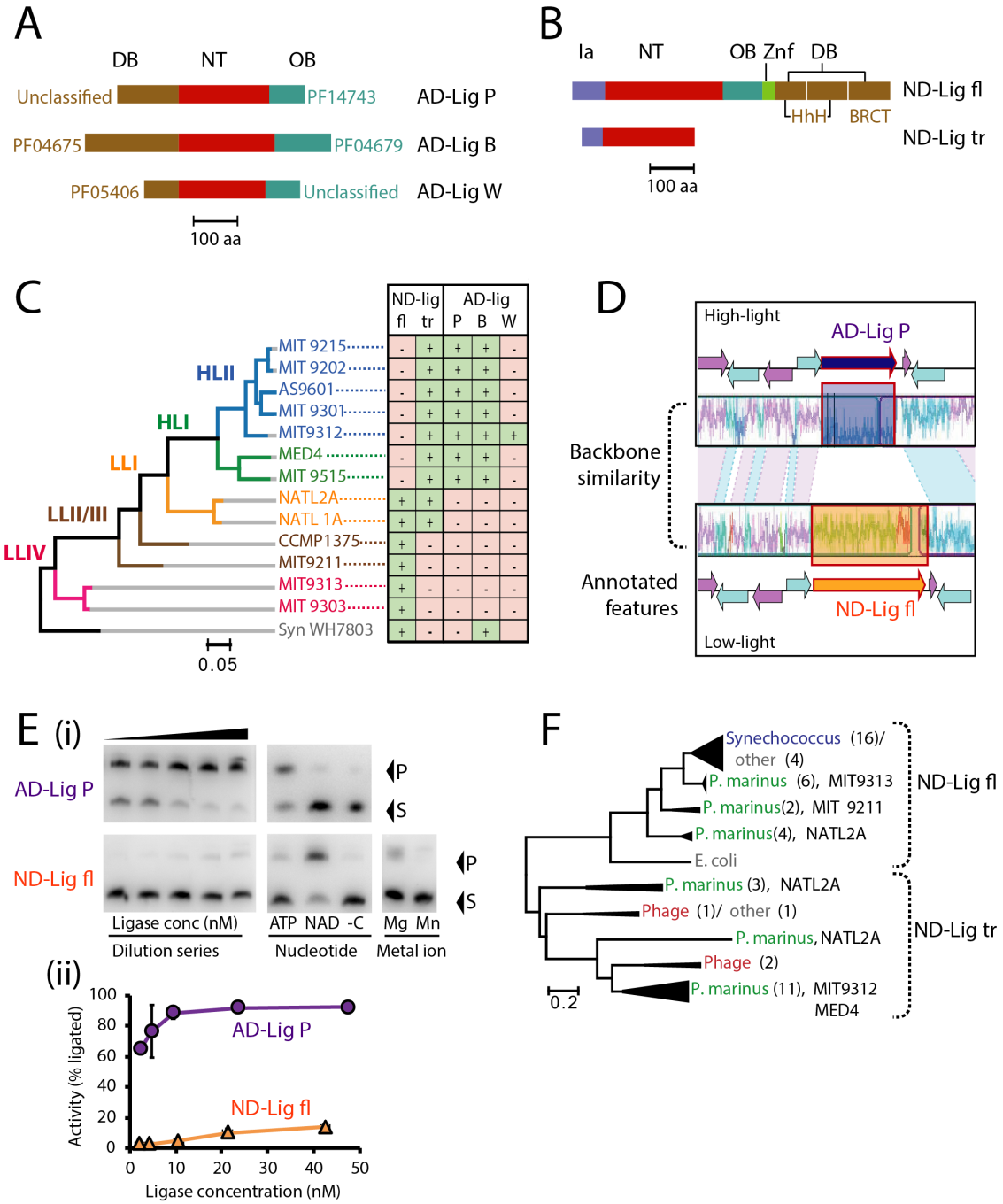

Figure 1. High-light and low-light *P.marinus* ecotypes have different complements of DNA ligases. A) Schematic of ATP-dependent DNA ligases found in high-light *P. marinus,* classified by Pfam assignment as described in (4). Domains are colored by function and Pfam identifiers are given adjacent. The structurally-conserved conserved catalytic core nucleotidyl transferase domain (NT) and oligonucleotide-binding (OB) domains are colored red and cyan; the variable DNA-binding domain (DB) is beige. B) Schematic of full-length (ND-Lig fl) and truncated (ND-lig tr) NAD-dependent DNA ligases colored by domain function. The OB, zinc-finger (ZnF) and DB domains are essential for activity in homologs (6). C) Presence and absence of ATP- and NAD-

dependent DNA ligases in the 13 closed genomes of *P. Marinus strains* with *Synechococcus* WH7803 as an outgroup. Data for all 41 closed and genomes and scaffolds is given in data S1. Strain evolution is inferred from the DNA gyrase B gene as described (15), and ecotypes are colored according to their assignment in ProPortal (11) with abbreviations high-light (HL) and low-light (LL). D) Genetic context of the ATP-dependent DNA ligase AD-Lig P and full-length NAD-dependent DNA ligase (ND-Lig fl) in high-light strain MIT9312 and low-light strain NATl2A respectively. Aligned regions of backbone similarity are colored violet/cyan and annotated genes are shown as arrows. No similarity was detected for the region encoding the ligase sequences AD-Lig P and ND-Lig fl. For entire alignment with representatives of all clades see data S2. E) DNA ligase activity for high-light AD-Lig P (MIT9302) and low-light ND-Lig fl (MIT9211). (i) Representative urea PAGE gels of ligation including specific activity by dilution (left), nucleotide cofactor preference (center) and metal ion preference (right). (ii) Ligation as a function of enzyme concentration integrated from left-most panel above. Data is the average of three replicates and error is the standard deviation of the mean. F) Maximum likelihood tree of representative NAD-dependent DNA ligase proteins from BLAST hits to MIT9313 ND-Lig fl and MIT9312 ND-Lig tr. All sequences were classified as full-length or truncated based on polypeptide length. All truncated sequences were less than 260 residues and included only Ia and NT domains. Sequences were trimmed to remove non-aligned portions of full-length ligases prior to tree building.

Supplementary materials for this manuscript include the following:

Datasets S1 to S2

Dataset S1. DNA ligases detected in 41 complete and partial genomes of *P. marinus* strains

Dataset S2. Whole genome alignment of representative strains from P. marinus ecotypes MIT_9312 (HLII), MED4 (HLI), NATL2A (LLI), MIT_9211 (LLII/III), MIT_9313 (LLIV). Genomes were aligned in progressiveMauve using default settings (Darling AE et.al (2010) *PLoS One* 5(6):e11147).